Primer

Prediction of protein structure

David Shortle

Proteins build their three-dimensional structures from the bottom up, utilizing bonding interactions between atoms in the backbone, the sidechains and water. And they do it with astonishing speed and efficiency. Although protein chemists would love to understand and model protein folding at this detailed level of physical chemistry, it is simply too formidable a challenge — now and for the foreseeable future. Not only is the abstract space of all allowed conformations truly astronomical in its size and complexity, it is extremely difficult to explore. Movement from one compact conformation to another is severely restricted by steric barriers, and the energy functions that provide a compass to find conformations lower in energy are inaccurate.

All general methods for predicting the structures of proteins from sequence that have met with some success can be viewed as proceeding in the opposite direction — from the top down. Starting with the databases of known protein structures and sequences, these methods employ a variety of tactics for recognizing patterns that connect sequence to three-dimensional structure, yet they can all be viewed as implementations of a common strategy.

In general terms, this strategy can be summarized as follows: the known structure of a single protein, a family of proteins, or many different proteins is represented symbolically in a structural template, a linear array of individual amino acid positions each of which corresponds to a specific structural environment (see Figure 1). Part or all of the amino acid sequence of the protein whose structure is to be predicted (known as the target) is inserted or aligned within the structural template, so that each residue occupies one position. The quality of the match or fit of each amino acid residue to the structural environment in which it finds itself is evaluated by calculating a score based on the observed frequency of occurrence of that amino acid type in similar structural environments. The more negative the score the worse the match and the more positive it is the better the match. These scores are then added together to give an overall score for the complete target sequence.

In some methods, the target sequence may be successively shifted within the template to generate a series of different alignments, each being given a separate score. In other methods, scoring is carried out on a set of different structural templates to find that template which yields the highest score. But the overall strategy remains the same — to identify the one structural template that gives the best overall score with the target sequence. Its structure forms the basis for all predictions about the structure of the target protein.

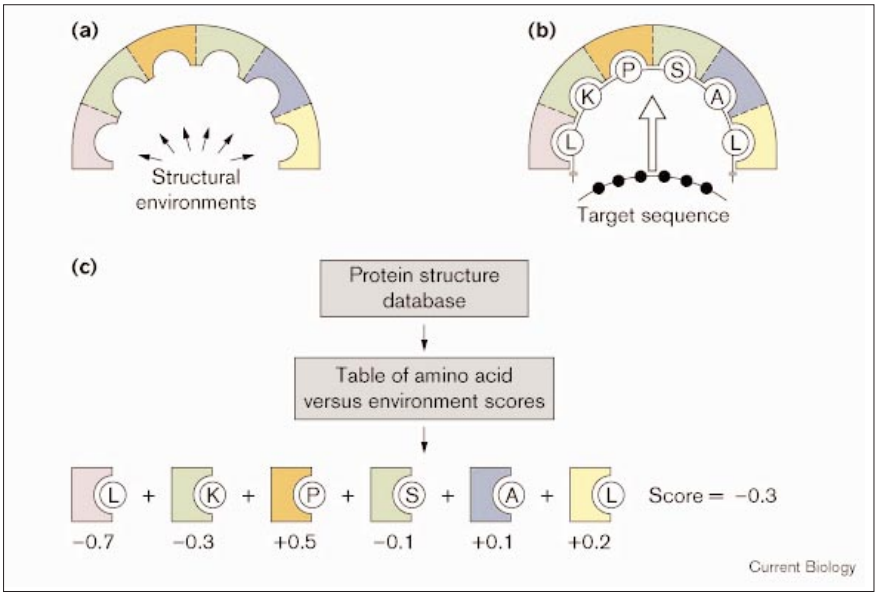
Homologues of known structures

Without question, the most reliable of all structure prediction methods is the familiar search for homology between the sequence of a protein and sequences of proteins of known structure. In this approach, the structural template is a protein's amino acid sequence, which acts as a surrogate for the residue positions in its three-dimensional structure. A variety of scoring functions can be employed based on sequence identity, various types of sequence similarity and penalties for introducing gaps into the sequence, yet they all permit estimation of the probability that the target and template sequences could have attained some level of similarity by random chance. When the score indicates this probability is vanishingly small, one can conclude that the two proteins have descended from a common ancestor and, with near certainty, will have similar three-dimensional structures for the chain segments they have in common.

Direct comparison of the target with a structural homologue is not the only way to identify an evolutionary relationship between them. Even when their sequences display no significant identity, sophisticated search algorithms may establish homology between very distantly related proteins through a series of bridging connections between the more closely related members of a large homologous family. As more sequences and three-dimensional structures are added to the databases, homology search methods become increasingly powerful. Although many protein families are not represented in the set of currently known folds, this situation will change dramatically in the next 10–20 years. The emerging field of structural genomics has as one of its primary goals the structural characterization of all fold families (see *Quick guide*, *Curr Biol* 1999, **9:**R871-R872).

Once a homologue of known structure has been identified, the goal of structure prediction shifts to obtaining a more detailed model of the target's structure. Although this will resemble its homologue, as the level of their shared sequence identity decreases, surface loops will vary increasingly in length and position, and larger numbers of substitutions within the protein interior will alter the angles between helices and strands and the packing of sidechain rotamers (alternative sidechain conformations formed by bond rotations). Much effort has been focused on correctly modeling

Figure 1



A generic strategy for predicting protein structure. **(a)** A structural template. **(b)** Alignment of the target sequence on the template. **(c)** Scoring the quality of the match of amino acid residues to the structural environments they occupy in the template. The scoring table is derived from a statistical analysis of the occurrence of each amino acid in each type of environment within proteins of known structure.

these changes in loops and sidechain rotamers, and a number of software packages are available that attempt to do this. Evaluation of these methods in three biannual meetings (the Critical Assessment of Structure Prediction, or CASP, meetings) has pointed to the difficulty of improving upon the structure of the closest homologue. The most salient conclusion to come from evaluating large numbers of blind predictions made with these methods is that the best models of the target incorporate the largest number of features directly from the structures of its homologues.

Predicting helices, strands and turns

If sequence searching fails to uncover a homologue of known structure, one can turn to a complicated battery of methods that predict a variety of types of structural information with some reliability. These methods differ primarily in the form of the structural template and the types of empirical information incorporated into the scoring function. A common place to start is the prediction of the target's secondary structure. Although the objectives of predicting a protein's structure usually involves more than finding α-helices, β-strands and turns, secondary structure prediction methods are simple to use and might provide clues as to which structural families the target might belong to.

The three types of secondary structure (α-helix, β-strand and neither) each have significantly different frequencies of the 20 amino acids. These patterns can be expressed in a variety of statistical forms, from single residue preferences or propensities to more complex correlations involving seven or more contiguous residues as a set (known as a window). The earliest methods for predicting secondary structure were carried out by hand and involved little more than averaging the secondary structure propensities of amino acids over a short window of several residues. The latest methods use structural templates constructed from complex non-linear algorithms known as neural nets, and hidden Markov models that encode the observed correlations between a residue at position i and its neighbors on either side out to i ± 3 or beyond.

With the best methods, residues in a particular sequence can be assigned to one of three structural categories — α-helix, β-strand, or neither — with average success rates of roughly 60–70%. This accuracy can be improved to almost 75% by repeating the process on the sequences of homologues of the target, insisting that the secondary structure must be the same for all of the members of the family. A related set of methods attempt to predict the percentage composition of α-helix, β-strand and irregular structure from the amino acid composition (without regard to sequence), permitting assignment of proteins to all α, all β and mixed α/β. Many of these methods are available for use online by submitting sequences to a remote server.

Fold recognition via 'threading'

Threading can be viewed as a direct extension of searching for sequence homology. Instead of using the amino acid sequence of a known structure, the structural template is the three-dimensional structure itself (or a simplified representation of it). During the search, the sequence is literally 'threaded' through the structural template. As with homology searches, gaps are allowed, especially in surface turns and loops, generating many different alignments. Although a variety of types of scoring function have been tried, the most common calculate an apparent 'energy' for each alignment based on the spatial distribution of the 210 types of amino acid pairs. From the database

of protein structures, the frequency of a given pair (for example, leucine–valine) separated by distances of 5–10 Å is tabulated. After normalization for chance proximity, the observed biases in these distributions of pairs are distilled down to sets of parameters known as empirical pair potentials. Several of the most distinctive features of protein structure, for example burial of hydrophobic residues and salt bridges located on the surface, are captured by these parameters.

Once the sequence has been aligned on the structure, the likelihood or probability of that three-dimensional distribution of residues is calculated as an energy term by summing over all pairs. Threading of a naturally occurring sequence through a set of hundreds of different 'decoy' structures almost always identifies the correct structure as the one with the lowest energy. Threading also recognizes structural homologues, with those of higher sequence similarity usually giving better scores than more distantly related homologues. When the level of sequence identity drops below the level of statistical significance, threading and the best methods of sequence searching provide complementary methods for fold recognition and display similar success rates. And like sequence searching, threading will become a more useful tool as the number of solved protein structures increases.

Ab initio prediction

When threading and sequence searching fail to identify a known fold that might resemble the target protein, an attempt can be made to predict the structure from 'first principles'. The traditional approach to *ab initio* prediction is to generate as many different conformations as possible and calculate the energy of each. When the search is terminated (computer resources are always finite), the conformation lowest in energy is deemed the predicted structure. As mentioned already, the vastness of conformational space and the technical difficulties of searching it efficiently have prevented this approach from being of much practical use. Consequently, current methods employ one or more short cuts to speed generation of a more diverse (but still quite small) sample of conformations. The two most popular tactics are to simplify the representation of the chain by fusing several atoms into 'united' atoms and to make conformation space discrete by either positioning atoms on a fixed three-dimensional lattice or constraining the allowed values of the dihedral angles ϕ, ψ and χ_1 to a small set.

The logic motivating these short cuts is that it may be possible to predict structure first at low resolution using simple models and then at higher resolution using more physically realistic models. Results to date have been disappointing. As succinctly stated by Einstein, "Models should be made as simple as possible, but no simpler." Unfortunately, it is not yet clear to what level computer models of protein chains and their interactions can be simplified without compromising the physical chemistry that determines structure.

Given the success of threading and empirical pair potentials in fold recognition, the current 'best bet' is that progress in *ab initio* prediction will involve small steps cantilevered out from the database of known structures. New conformations can be constructed by combining pieces of structure taken from real proteins (otherwise known as Frankenstein conformations), thereby saving computer time otherwise wasted on grossly unrealistic conformations. The most noteworthy successes in the *ab initio* category at the most recent CASP meeting (1998) were based on this approach.

Structure prediction is still at the stage of a developing empirical technology, based on rules of thumb that work in certain situations, yet lack a quantitative scientific explanation. Consequently, beyond recognition of the correct fold by identifying a structural homologue, prediction of the details of protein structure remains an uncertain, probabilistic enterprise. Until the physical chemistry underlying protein structure can be modeled more accurately by methods that proceed from the bottom up, caveat predictor!

Key references

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402

Bowie JU, Eisenberg D: **Inverted protein structure prediction**. *Curr Opin Struct Biol* 1993, **3:**437-444.

Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins* 1997, **1 (suppl).**

Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins* 1999, **3 (suppl).**

Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition**. *Nature* 1992, **358:**86-89.

Novotny J, Sippl M: **Theory and simulation: old problems, new paradigms**. *Curr Opin Struct Biol* 1997, **7:**179–180.

Address: The Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, Maryland 21205, USA.
E-mail: shortle@welchlink.welch.jhu.edu